Simões, Barreiro, Santos, Sousa-Silva & Tagnin (eds.) *Linguística, Informática e Tradução: Mundos que se Cruzam*, Oslo Studies in Language 7(1), 2015. 397–424. (ISSN 1890-9639 / ISBN 978-82-91398-12-9)
http://www.journals.uio.no/osla

# AS WORDNETS DO PORTUGUÊS

HUGO GONÇALO OLIVEIRA, VALERIA DE PAIVA, CLÁUDIA FREITAS, ALEXANDRE RADEMAKER, LIVY REAL E ALBERTO SIMÕES

ABSTRACT

Not many years ago it was usual to comment on the lack of an open lexical-semantic knowledge base, following the lines of Princeton WordNet, but for Portuguese. Today, the landscape has changed significantly, and researchers that need access to this specific kind of resource have not one, but several alternatives to choose from. The present article describes the wordnet-like resources currently available for Portuguese. It provides some context on their origin, creation approach, size and license for utilization. Apart from being an obvious starting point for those looking for a computational resource with information on the meaning of Portuguese words, this article describes the resources available, compares them and lists some plans for future work, sketching ideas for potential collaboration between the projects described.

## [1] INTRODUÇÃO

Relações semânticas são um aspecto fundamental a ser levado em conta quando se pretende construir programas de computador capazes de lidar com o conteúdo de textos -- elas estabelecem associações de sentido entre palavras e podem ser integradas em bases de conhecimento léxico-semântico, como a WordNet de Princeton (Miller 1995; Fellbaum 1998, 2010). Disponível desde o início da década de 1990, a WordNet de Princeton (doravante, WN.Pr) é um recurso paradigmático: embora criada apenas para o inglês, seu modelo é quase um *standard*, o que se comprova pela sua ampla utilização e adaptação a diferentes línguas (Bond & Paik 2012).

Quanto à língua portuguesa, só na década de 2000 foi anunciada a WordNet.PT. No entanto, e diferentemente da WN.Pr, esta nunca foi de livre utilização, o que, na prática, significou a continuação de uma lacuna para o português. Por outro lado, e paralelamente, surgiram algumas alternativas ao modelo de wordnet, algumas delas alvo de uma comparação feita por Santos et al. (2010), que também aponta questões relacionadas à própria construção de wordnets.

Mas, se as alternativas existentes se mostraram proveitosas em algumas tarefas do processamento computacional da língua portuguesa — veja-se, por exemplo, a utilidade da redes de palavras para o português (Gonçalo Oliveira 2014) —

continuava a faltar uma wordnet propriamente dita para esta língua, que tornasse possível a utilização de abordagens usuais no processamento de linguagem natural (PLN), tais como o cálculo de similaridade (Resnik 1995) ou a desambiguação do sentido das palavras (Banerjee & Pedersen 2002), dependentes precisamente da existência de uma wordnet para a língua alvo.

Foi neste contexto que, no início da década de 2010, surgiram não um, mas vários projetos que disponibilizaram gratuitamente wordnets para esta língua, criados em diferentes contextos e seguindo diferentes abordagens.

Este artigo, escrito pelos responsáveis por três desses projetos, descreve as várias wordnets que existem atualmente para a língua portuguesa, indicando o contexto em que foram criadas, o processo de construção, a sua disponibilização e, dentro do possível, a sua dimensão. O artigo pode ser visto como uma continuação de Santos et al. (2010), ainda que focado essencialmente em recursos que adotaram o modelo original da WN.Pr.

Na secção [2] é feita precisamente uma breve apresentação da WN.Pr, com uma referência ao seu modelo, à sua adaptação a outras línguas e à sua expansão através do alinhamento de conteúdos. Seguem-se várias descrições das wordnets do português, começando por aquelas que não estão disponíveis gratuitamente (secção [3]), passando depois para outros recursos léxico-semânticos, todos eles relacionados com as wordnets e a certa altura utilizados como alternativa à WN.Pr (secção [4]), e finalizando com as wordnets livres do português (secção [5]). A secção [6] traz uma visão comparativa, onde as várias wordnets e um conjunto de algumas das suas propriedades qualitativas e quantitativas são colocadas lado a lado. Para concluir, apresentamos na secção [7] algum trabalho futuro planeado para as wordnets de que os autores deste artigo são responsáveis, seguido imediatamente de algumas ideias de colaboração que, acreditamos, serão importantes para estabelecimento destes recursos como alternativas de qualidade para o processamento computacional da língua portuguesa.

## [2] O MODELO WORDNET

Bases de conhecimento lexical são repositórios organizados de itens lexicais. Entre outras informações, estes recursos incluem normalmente informação sobre os possíveis sentidos das palavras, relações entre sentidos, definições e frases que exemplificam a sua utilização. O modelo da wordnet, criado para a WN.Pr tendo o inglês como língua alvo, é provavelmente o modelo mais popular para representar este tipo de recurso. Sua flexibilidade levou não só à crescente aceitação por parte da comunidade PLN, mas também à sua adaptação para outras línguas, tornando-se quase um *standard*.

[2.1] *WordNet de Princeton: a mãe de todas as wordnets*

A WN.Pr foi criada manualmente no início da década de 1990, e vem sendo atualizada desde então. Inicialmente baseada em princípios psicolinguísticos, combina informação lexicográfica tradicional, semelhante à encontrada num dicionário, com uma organização adequada para a utilização computacional, o que facilita a sua utilização como base de conhecimento léxico-semântico.

Tal como num tesauro, a WN.Pr é organizada em grupos de itens lexicais sinónimos, chamados de *synsets*, que podem ser vistos como as possíveis lexicalizações para um conceito de uma língua. Além da relação de sinonímia, inerente aos *synsets*, a WN.Pr abrange outros tipos de relação semântica, estabelecidos entre os *synsets*, para além de algumas relações entre itens lexicais. Entre as relações semânticas abrangidas, temos, por exemplo, a hiperonímia — o conceito representado por um *synset* é uma generalização de outro — e a meronímia — o conceito representado por um *synset* é uma parte de outro.

Para além desta informação semântica, cada *synset* pertence a uma determinada categoria gramatical (substantivo, verbo, adjetivo ou advérbio); tem uma glosa, semelhante a uma definição num dicionário; e pode ter ainda frases que ilustram o emprego de algumas das suas palavras. A inclusão de um item lexical num *synset* indica um sentido desse item. A figura 1 mostra, para a palavra *bird*, os *synsets* na WN.Pr 3.0. Para esta palavra, estão definidos cinco sentidos nominais e um verbal. Para cada *synset*, apresenta-se a sua glosa (entre parênteses) e expandiu-se a lista de hipónimos diretos do primeiro *synset*. Sobre a ordem de apresentação dos *synsets* e dos itens que incluem, há a dizer que, sempre que possível, são consideradas as respectivas frequências no corpo SemCor (George A. Miller and Martin Chodorow and Shari Landes and Claudia Leacock and Robert G. Thomas 1994), onde esta informação se encontra manualmente anotada.

Apesar de algumas críticas ao modelo da WN.Pr (Sampson 2000), este é sem dúvida um recurso muito completo, especialmente se considerarmos que foi criado manualmente. Outros pontos importantes para o seu sucesso e ampla utilização foram, por um lado, a flexibilidade do seu modelo e, por outro, a sua disponibilização gratuita. O primeiro tornou possível a integração da WN.Pr numa grande quantidade de projetos de PLN ou de gestão de conhecimento, tornando o modelo WN.Pr praticamente *standard* com relação a várias línguas; o segundo fez com que isso fosse possível sem quaisquer custos monetários.

A crescente popularidade deste modelo de base de conhecimento levou à criação da Global WordNet Association (GWA), uma organização não comercial que oferece uma plataforma para a discussão, partilha e ligação das wordnets no mundo. Para um levantamento de wordnets e suas licenças, ver (Bond & Paik 2012), ou a lista, mais atualizada, disponível a partir da página da GWA.[1]

[1] Ver http://globalwordnet.org/wordnets-in-the-world/

**Noun**

- bird (warm-blooded egg-laying vertebrates characterized by feathers and forelimbs modified as wings)
  **[direct hyponym]**
  - dickeybird, dickey-bird, dickybird, dicky-bird (small bird; adults talking to children sometimes use these words to refer to small birds)
  - cock (adult male bird)
  - hen (adult female bird)
  - nester (a bird that has built (or is building) a nest)
  - night bird (any bird associated with night: owl; nightingale; nighthawk; etc)
  - parrot (usually brightly colored zygodactyl tropical birds with short hooked beaks and the ability to mimic sounds)
  - ...
- bird, fowl (the flesh of a bird or fowl (wild or domestic) used as food)
- dame, doll, wench, skirt, chick, bird (informal terms for a (young) woman)
- boo, hoot, Bronx cheer, hiss, raspberry, razzing, razz, snort, bird (a cry or noise made to express displeasure or contempt)
- shuttlecock, bird, birdie, shuttle (badminton equipment consisting of a ball of cork or rubber with a crown of feathers)

**Verb**

- bird, birdwatch (watch and study birds in their natural habitat)

FIGURA 1: *Synsets* com a palavra *bird* na WordNet de Princeton 3.0 e os primeiros hipónimos a partir de seu primeiro significado.

### [2.2] *WordNets multilíngues e outros alinhamentos*

No contexto da GWA, foi também estudada a possibilidade de alinhar, dentro do possível, wordnets de diferentes línguas, dadas as suas semelhanças. Assim, começaram a surgir algumas wordnets multilíngues, como a EuroWordNet (Vossen 1997) ou a MultiWordNet (Pianta et al. 2002), ainda que seguindo abordagens diferentes de desenvolvimento. Na EuroWordNet, wordnets são criadas de forma independente para cada língua, procurando-se depois alinhar semelhanças entre elas ou, indiretamente, através da WN.Pr, com recurso ao chamado *Inter-language Index*. Na MultiWordNet, o primeiro passo é traduzir, dentro do possível, uma wordnet "pivot", normalmente a WN.Pr, o que garante algum alinhamento.

Entre outras wordnets multilíngues, também alinhadas à WN.Pr, destacam-se a BalkaNet (Stamou et al. 2002), dedicada às línguas dos Balcãs, e o Multilingual Central Repository (Gonzalez-Agirre & Rigau 2013) (doravante, MCR), dedicado às línguas faladas em Espanha.

A Open Multilingual WordNet (Bond & Foster 2013) (doravante, OMWN) é uma iniciativa que visa facilitar o acesso a diferentes wordnets, para diferentes línguas. Para tal, wordnets criadas de forma independente, foram normalizadas, ligadas à WN.Pr e tornadas acessíveis através de uma interface comum.[2]

Outra iniciativa que deve ser mencionada é a Universal WordNet (de Melo & Weikum 2009) (doravante, UWN), uma base de conhecimento lexical multilíngue construída automaticamente com base na WN.Pr e no alinhamento de versões multilíngues da Wikipédia, desenvolvida no Instituto de Informática Max Planck, na Alemanha. A UWN estende a WN.Pr com cerca de 1,5 milhões de ligações de significado (*meaning links*) para 800 mil palavras em mais de 200 línguas, apresentando evidência extraída a partir de uma variedade de meios incluindo wordnets (monolíngues) pré-existentes, dicionários bilíngues e corpos paralelos alinhados.

Há também vários alinhamentos entre WN.Pr e outros recursos, incluindo as ontologias SUMO (Pease & Fellbaum 2010) e DOLCE (Gangemi et al. 2010), e bases de conhecimento que integram a WN.Pr com outros recursos como a Wikipédia, onde se destaca o YAGO (Suchanek et al. 2007); a Wikipédia e outros recursos léxico-semânticos, onde se destacam a UBY (Gurevych et al. 2012), ou a BabelNet (Navigli & Ponzetto 2012). Por exemplo, a BabelNet, atualmente na versão 3.0, abrange 271 línguas, incluindo o português, o que é possível através do alinhamento da WN.Pr com as versões da Wikipédia para várias línguas, a que se junta ainda informação do Wikcionário,[3] OmegaWiki,[4] Wikidata[5], e das wordnets que fazem parte da OMWN (Bond & Foster 2013).

[2] Ver http://compling.hss.ntu.edu.sg/omw/
[3] Ver https://www.wiktionary.org/
[4] Ver http://www.omegawiki.org/
[5] Ver http://www.wikidata.org/

## [3] WORDNETS FECHADAS DO PORTUGUÊS

Não há dúvidas que, para além da flexibilidade do seu modelo, o caráter de domínio público da WN.Pr foi um fator chave na sua aceitação. Apesar disso, nem todos os recursos que seguem este modelo optaram por tornar o seu resultado livre. Neste leque encontra-se a WordNet.PT, aquela que foi a primeira wordnet do português, mas que se encontra disponível apenas para exploração através da sua página web, não sendo possível ser descarregada para utilização local ou integração em diferentes projetos. Para além da WordNet.PT, esta secção descreve outros dois projetos que resultaram na criação de uma wordnet para o português e que, por alguma razão, não se encontram disponíveis ou, pelo menos, disponíveis gratuitamente. São eles a WordNet.BR, um projeto, aparentemente, inacabado, e para o qual apenas estão disponíveis os *synsets*, sob a forma do thesaurus eletrónico TeP; e a MWN.PT que pode ser explorada tanto através da sua página web como da página do projeto MultiWordNet, mas só pode ser descarregada mediante o pagamento de uma licença académica ou comercial.

### [3.1] *WordNet.PT*

A WordNet.PT (Marrafa 2001, 2002) (doravante, WN.PT) terá sido a primeira wordnet para o português. Desenvolvida desde 1998, é um projeto coordenado por Palmira Marrafa, no Centro de Linguística da Universidade de Lisboa, mais propriamente no CLG — Grupo de Computação do Conhecimento Léxico-Gramatical, em colaboração com o Instituto Camões.

A sua construção é essencialmente manual e segue o modelo da EuroWordNet (Vossen 1997), ou seja, a WN.PT é criada de raíz para a língua portuguesa. A sua versão mais recente, WN.PT 1.6, data de 2006 e abrange várias relações semânticas, nomeadamente: geral/específico (incluindo *hiperonímia*), todo/parte, equivalência, oposição, categorização, e ainda relações entre os participantes num evento (incluindo *instrumento-para* ou *lugar-para*) e definidoras da estrutura de um evento (incluindo *estar-envolvido-em* ou *lugar-para*). A mesma versão cobre os seguintes domínios semânticos: atividades artísticas e profissionais, comida, regiões geográficas e políticas, instituições, instrumentos, meios de transporte, vias de comunicação, obras de arte, saúde e atos médicos, seres vivos e vestuário.

Mais recentemente, este recurso foi expandido para WordNet.PT Global — *Rede Léxico-Conceptual das variedades do Português* (Marrafa et al. 2011), que pretende incluir variantes de outros países de língua oficial portuguesa. De acordo com a informação na sua página web,[6] a WN.PT Global contém uma rede de 10 mil conceitos, incluindo substantivos, verbos e adjetivos, as suas lexicalizações nas diferentes variantes do português e as suas glosas. Os conceitos estão integrados em uma rede com mais de 40 mil instâncias de relação. Em 2014, foi apresentada uma

[6] Ver http://cvc.instituto-camoes.pt/traduzir/wordnet.html

primeira abordagem para expandir a WN.PT de forma semi-automática (Amaro 2014), através da extração de relações a partir de um corpo, o que mostra que, ainda que fechado, este projeto continua ativo.

### [3.2] *WordNet.Br*

A WordNet.BR (Dias-da-Silva et al. 2002; Dias-da-Silva 2006) (doravante, WN.BR) foi desenvolvida sob a coordenação de Bento Dias da Silva, na Faculdade de Ciências e Letras da Universidade Estadual Paulista, com vista a criar uma wordnet para a variante brasileira do português. Numa primeira fase de desenvolvimento (Dias-da-Silva et al. 2002), uma equipa de três linguistas analisou cinco dicionários de português do Brasil e dois corpos, de forma a obter informação sobre sinonímia e antonímia. Esta fase resultou na criação manual de *synsets* e relações de antonímia entre eles, bem como na escrita de algumas glosas e seleção de frases exemplo.

Numa segunda fase, os *synsets* da WN.BR foram alinhados manualmente com a WN.Pr (Dias-da-Silva 2006), num processo semelhante ao seguido no projeto EuroWordNet, onde se recorreu a dicionários bilíngues. Após o alinhamento com a WN.Pr, as relações semânticas estabelecidas entre *synsets* com equivalências em português e inglês foram herdadas.

Com base no processo relatado, supõe-se que a versão completa da WN.BR cobrirá as relações de hiperonímia, parte-de, causa e implicação (*entailment*). No entanto, esta versão não se encontra disponível na rede, provavelmente por a segunda fase de desenvolvimento não ter sido concluída. Por outro lado, é possível consultar e descarregar os resultados da primeira fase, disponíveis sob o nome de TeP (Maziero et al. 2008) — *Thesaurus Eletrônico do Português*. O TeP é mantido pelo Núcleo Interinstitucional de Linguística Computacional (NILC) da Universidade de São Paulo, em São Carlos, Brasil. Inclui mais de 44 mil itens lexicais, organizados em 19.888 *synsets*, que por sua vez estão ligados através de 4.276 relações de antonímia.

### [3.3] *MultiWordNet.PT*

A MultiWordNet.PT, normalmente referida como MWN.PT,[7] é a parte portuguesa do projeto MultiWordNet (Pianta et al. 2002). Foi desenvolvida pelo NLX - *Natural Language and Speech Group*, na Universidade de Lisboa, e pode ser comprada através do catálogo da European Language Resources Association.[8]

De acordo com a sua documentação,[9] a MWN.PT inclui 17,2 mil *synsets* validados manualmente, o que corresponde aproximadamente a 21 mil sentidos e 16 mil lemas, que abrangem tanto a variante europeia como a variante brasileira

[7] Ver http://mwnpt.di.fc.ul.pt/
[8] Ver http://catalog.elra.info/
[9] Ver http://mwnpt.di.fc.ul.pt/features.html

do português. Sendo um recurso criado no âmbito do projeto MultiWordNet, os *synsets* da MWN.PT derivam da tradução dos seus equivalentes na WN.Pr, recurso com que a MWN.PT está alinhada. Transitivamente, este recurso acaba por estar também alinhado com as MultiWordNets do italiano, espanhol, hebreu, romeno e latim.

Os *synsets* da MWN.PT estão ligados através das relações de hiperonímia/hiponímia e meronímia (parte, membro e substância), e este recurso inclui as subontologias sob os conceitos de pessoa, organização, evento, localização, e obras de arte. Alegadamente, este recurso cobre os 98 conceitos base sugeridos pela GWA e os 164 conceitos nucleares (*core base concepts*) indicados pela EuroWordNet como aqueles que estão presentes em todas as wordnets do projeto.[10]

Em Santos et al. (2010) verificou-se que a MWN.PT não tinha informação sobre a categoria gramatical dos *synsets*. Após realizarmos várias buscas a este recurso não encontramos nenhum resultado para palavras de outras classes que não fossem substantivos, nem mesmo para verbos frequentes como "faltar" ou "ter". Por isso, admitimos que todas as palavras são substantivos. No entanto, os conceitos nucleares incluem não só 129 substantivos (66 concretos e 63 abstratos), mas também 35 verbos abstratos, que não estarão abrangidos. Para além disso, não foi possível encontrar correspondências equivalentes em português para alguns dos *synsets* abrangidos pela MWN.PT. Por exemplo, os conceitos nucleares *human_action* e *magnitude_relation* da WN.Pr estão alinhados com um `GAP!` na MWN.PT. A presença de `GAP!` ou `PSEUDOGAP!`, interpretados como falhas lexicais, realça precisamente uma limitação da tradução direta de uma wordnet "pivot" para uma outra língua.[11] No entanto, devido à escassa documentação deste recurso (Santos et al. 2010), não temos como garantir que as *gaps* sejam fruto realmente de lacunas lexicais, uma vez que não sabemos quais critérios nortearam a tradução da WN.Pr neste projeto.

## [4] ANTES DAS WORDNETS LIVRES

A criação manual de uma wordnet é uma tarefa complexa e que requer muito tempo. Assim, durante a década de 2000, investigadores da área do PLN em português que necessitavam e não tinham acesso à WordNet.PT tiveram de encontrar alternativas livres, que, na maior parte das vezes, eram também mais simples.

Neste âmbito, para além do TeP (Maziero et al. 2008), já mencionado na secção [3.2], destacam-se:

[10] Mais sobre estas listas de conceitos pode ser consultado em `http://www.globalwordnet.org/gwa/ewn_to_bc/topont.htm`

[11] Por outro lado, a estratégia da tradução é sempre válida quando a alternativa é a ausência de recurso ou de recurso alinhado, considerando-se que o interesse está na tradução.

- O OpenThesaurus.PT,[12] versão portuguesa correspondente ao projeto homónimo, OpenThesaurus (Naber 2004), normalmente utilizado para sugerir sinónimos em processadores de texto;

- O PAPEL (Gonçalo Oliveira et al. 2008), uma rede extraída automaticamente a partir de um dicionário da língua portuguesa, e que liga palavras relacionadas por um vasto leque de relações. Mais recentemente, o PAPEL foi expandido para CARTÃO (Gonçalo Oliveira et al. 2011), com base na exploração de mais dicionários;

- Alguns dos recursos desenvolvidos no âmbito do Port4Nooj (Barreiro 2010), construídos no ambiente de desenvolvimento linguístico do NooJ (Silberztein 2005), inicialmente extraídos do sistema de tradução automática OpenLogos (Barreiro et al. 2014). Estes recursos incluem, por exemplo, um conjunto de definições e relações semânticas entre palavras;

- O Dicionário Aberto (Simões et al. 2012), no qual, juntamente com um dicionário, são disponibilizadas relações entre as suas palavras.

Uma descrição mais pormenorizada destes recursos, alguns dos quais comparados em Santos et al. (2010), está contudo fora do âmbito deste artigo.

## [5] WORDNETS LIVRES DO PORTUGUÊS

No início da década de 2010 surgiram várias wordnets para português. Todas elas têm a particularidade de terem sido criadas de forma automática ou semi-automática, ainda que seguindo metodologias diferentes, e de partirem do princípio que recursos léxico-semânticos precisam ser abertos para serem realmente úteis à comunidade. Esta secção apresenta, por ordem cronológica do seu primeiro anúncio, as quatro wordnets que se enquadram nesta descrição e que, por isso, estão disponíveis gratuitamente na rede.

### [5.1] *Onto.PT*

A Onto.PT (apresentada inicialmente em (Gonçalo Oliveira & Gomes 2010), descrita de forma resumida em (Gonçalo Oliveira & Gomes 2014a), e detalhada em (Gonçalo Oliveira 2013)) é uma wordnet desenvolvida no âmbito do doutoramento de Hugo Gonçalo Oliveira, sob a orientação de Paulo Gomes, no Centro de Informática e Sistemas da Universidade de Coimbra. O projeto teve início nos finais de 2008, num contexto em que não existia uma wordnet livre para o português, nem recursos humanos para criar uma nova wordnet para esta língua. O objetivo foi sempre criar uma wordnet de forma completamente automática, aproveitando

[12] Até recentemente disponível a partir de `http://openthesaurus.caixamagica.pt/`

ao máximo os recursos desenvolvidos no âmbito do projeto PAPEL (Gonçalo Oliveira et al. 2008), nomeadamente gramáticas para extração de relações a partir de dicionários e a definição das relações extraídas. Ao mesmo tempo, tentou-se aproveitar outros recursos lexicais livres para o português, nomeadamente o Wiktionário.PT,[13] o Dicionário Aberto (Simões et al. 2012), o TeP (Maziero et al. 2008), o OpenThesaurus.PT e, mais recentemente, a OpenWN-PT (de Paiva et al. 2012; Rademaker et al. 2014).

A abordagem de construção da Onto.PT, apelidada de ECO (Gonçalo Oliveira & Gomes 2014a), é, no entanto, suficientemente flexível para integrar palavras e relações obtidas de outros recursos, o que poderá vir a ser feito no futuro. Ela distingue-se de abordagens baseadas em tradução: em alternativa a encontrar correspondência, em português, de palavras e *synsets* em wordnets de outras línguas, ECO tenta aprender automaticamente toda a estrutura de uma wordnet, incluindo os conteúdos e próprios limites dos *synsets*, ou os *synsets* envolvidos em cada instância de relação. Daí, e apesar de explorar, de forma automática, alguns recursos criados manualmente, os autores se referirem a ela como uma abordagem "completamente automática". A abordagem ECO é composta por três fases principais:

(i) Extração de relações entre palavras, o que até à data tem sido feito a partir de definições de dicionários.

(ii) Descoberta de aglomerados de palavras (*clusters*), através da exploração do grafo de relações de sinonímia. Esta fase pode ou não ter como ponto de partida um conjunto inicial de *synsets* já definido, como o do TeP.

(iii) Mapeamento de relações entre palavras em relações entre os *synsets* descobertos.

A figura 2 exemplifica estes três passos. Na sua versão mais recente, Onto.PT 0.6 (Gonçalo Oliveira & Gomes 2014b), há ainda uma fase em que definições de dicionário são associadas automaticamente a *synsets*.

A Onto.PT pode ser vista como uma wordnet um pouco diferente do normal. Isto verifica-se não só na abordagem de construção seguida, mas também por ser um recurso que inclui um vasto conjunto de relações semânticas, precisamente o mesmo do projeto PAPEL. Inclui assim não só as relações mais comuns, como a hiperonímia e vários tipos de meronímia, mas também outras relações como causa, finalidade, local ou maneira.

[13] Ver https://pt.wiktionary.org/

| | | | |
|---|---|---|---|
| | | **Extração** | |
| gado | s.m. | = | conjunto de animais criados para diversos fins; rebanho |
| | $triplo\_1$ | = | rebanho SINONIMO_DE gado |
| | $triplo\_2$ | = | animal MEMBRO_DE gado |
| | | **Clustering** | |
| | $synset_1$ | = | {manada, rebanho, mancheia, boiada} |
| $synset_1$ | $+tb - triple_1$ | = | {manada, rebanho, mancheia, boiada, gado} |
| | | **Mapeamento** | |
| | $synset_2$ | = | {bicho, animal, alimal, béstia, minante} |
| | $triplo_s yn_1$ | = | $synset_2$ MEMBRO_DE $synset_1$ |

FIGURA 2: Exemplo das três primeiras fases da abordagem ECO.

Por um lado, a abordagem ECO permite obter uma wordnet de grandes dimensões com pouco esforço — a versão 0.6 inclui cerca de 169 mil itens lexicais únicos, organizados em cerca de 117 mil *synsets*, que por sua vez se relacionam através de cerca de 174 mil instâncias de relação. Por outro, há consequências a nível da qualidade dos conteúdos. Por exemplo, na versão 0.35 do recurso, estimou-se que cerca de 74% dos *synsets* estavam corretos, em 18% não havia concordância entre avaliadores e os restantes tinham pelo menos uma palavra que não lhes devia pertencer (avaliação descrita de forma detalhada em (Gonçalo Oliveira 2013)). A qualidade das relações também varia drasticamente consoante o seu tipo. Considerando que relações entre *synsets* errados estão também erradas, as relações de hiperonímia estavam cerca de 65% corretas, número que aumentava para 78% a 82% num conjunto que incluía os restantes tipos de relação. Ainda assim, entre outras tarefas, a Onto.PT foi já usada na expansão de sinónimos para recuperação de informação (Rodrigues et al. 2012) ou de criação de listas de verbos causais (Drury et al. 2014).

Devido à sua abordagem de construção, a Onto.PT não é um recurso estático e pode, de versão para versão, ter mudanças significativas ao nível do número e tamanho dos *synsets*. Assim, no entender dos seus autores, não fará sentido tentar alinhá-lo com a WN.Pr. Há a acrescentar que a Onto.PT se encontra disponível gratuitamente[14] sob a forma de um modelo RDF/OWL, inspirado num modelo existente para representar a WN.Pr (van Assem et al. 2006), mas expandido para abranger outros tipos de relação.

[14] Ver http://ontopt.dei.uc.pt/

### [5.2] *OpenWordNet-PT*

A OpenWordNet-PT (de Paiva et al. 2012; Rademaker et al. 2014), abreviada como OpenWN-PT, é uma wordnet desenvolvida originalmente por Valeria de Paiva, Alexandre Rademaker e Gerard de Melo como uma projeção sintática da Universal WordNet[15] (UNW).

A OpenWN-PT está sendo desenvolvida desde 2010 com o objetivo principal de servir como subsídio léxico para um sistema voltado para raciocínio lógico, seja este desenvolvido usando lógicas descritivas (em processo de adaptação) ou lógicas de primeira-ordem, baseadas em representação do conhecimento, por exemplo usando a ontologia SUMO (Pease & Fellbaum 2010).

O processo de construção da OpenWN-PT, decorrente do processo de criação da UWN, usa aprendizagem de máquina para construir relações entre grafos que representam informação vinda de versões em múltiplas línguas da Wikipédia, bem como de dicionários eletrônicos abertos. Apesar de ter começado como uma projeção apenas ao nível dos lemas em português e suas relações, a OpenWN-PT tem sido constantemente melhorada por meio de acréscimos linguisticamente motivados, quer manualmente, quer fazendo uso de grandes corpos, como é o caso do léxico de nominalizações que integra a OpenWN-PT (de Paiva et al. 2014b; Freitas et al. 2014a). Uma das características da construção deste último recurso é tentar incorporar os diferentes materiais (de qualidade) já produzidos e disponibilizados para a língua portuguesa, independente de variante.

A OpenWN-PT integra três estratégias linguísticas no seu processo de enriquecimento lexical: (i) tradução; (ii) corpo; (iii) dicionários. Com relação à tradução, são usados léxicos e listas produzidas para outras línguas, como inglês, francês e espanhol, automaticamente traduzidos e posteriormente revistos. A incorporação de dados de corpos contribui com palavras ou expressões de uso corrente que podem ser específicas da língua portuguesa ou que, por outros motivos, podem não constar nas outras wordnets.

Como a Onto.PT, a OpenWN-PT também está disponível em RDF/OWL, seguindo e expandindo, quando necessário, o mapeamento proposto por van Assem et al. (2006). Tanto os dados da OpenWN-PT quanto as definições do modelo RDF (classes e propriedades) estão livremente disponíveis para *download*.[16] A filosofia da OpenWN-PT consiste em manter a ligação estreita com a WN.Pr, mas tentar remover os erros maiores criados pelos métodos automáticos, usando conhecimentos linguísticos. Uma consequência desta ligação estreita com a WN-Pr é a possibilidade de minimizar os impactos decorrentes de decisões lexicográficas quanto à

[15] Por projeção sintática, entenda-se uma projeção usando simplesmente a informação sintáctica de que registros correspondem a entradas em português, sem levar em conta o significado semântico do registro. Como esses registros são construídos automaticamente, pode haver casos em que a configuração foi equivocada, onde o processo automático de unificação decidiu que uma palavra em catalão era português, por exemplo.

[16] Ver https://github.com/arademaker/openWordnet-PT

separação ou agrupamento de sentidos em um *synset*. Como, em última análise, tais decisões serão sempre arbitrárias (Kilgarriff 1997), o critério prático do alinhamento multilíngue atua como uma solução bem vinda.

A OpenWN-PT foi escolhida pelos organizadores dos projetos FreeLing (Padró & Stanilovsky 2012), OMWN (Bond & Foster 2013) e ainda *Google Translate*[17] como a representante das wordnets abertas em português utilizada por esses projetos, respectivamente. Presumivelmente essa escolha se deve à cobertura abrangente da OpenWN-PT e também à sua qualidade. Embora os autores do recurso não tenham feito medições dessa qualidade, a UWN original produziu estatísticas impressionantes em termos de sua abrangência e precisão de seus dados — mais de 200 línguas, 1.595.763 ligações entre termos e significados, 822.212 termos, com precisão avaliada de mais de 89% em francês, mais de 85% em alemão e mais de 90% em chinês (mandarino) como descrito em (de Melo & Weikum 2012). A rede OpenWordNet-PT tem sido, depois de sua versão inicial baseada na UWN, constantemente revisada e aprimorada manualmente.

A OpenWN-PT tem no momento 43.925 *synsets*, dos quais 32.696 correspondem a substantivos, 4.675 a verbos, 5.575 a adjetivos e 979 a advérbios. Para além de descarregados, os dados podem ser consultados via SPARQL no respetivo *endpoint*[18] e a base pode ser consultada e comparada com outras wordnets,[19] usando o menu para trocar a língua de inglês para português.

### [5.3] *Ufes WordNet*

A Ufes[20] WordNet[21] (doravante, UfesWN.BR) é um projeto que visa a construção de um banco de dados léxico em Português do Brasil com estrutura similar à da WN.Pr (Gomes et al. 2013), baseando-se na tradução automática da WN.Pr. Para a tradução, foi construída uma ferramenta baseada na API do *Google Translate* especificamente com este propósito, e recorrendo ainda à biblioteca de acesso à WN.Pr, JWI (Finlayson 2014).

De acordo com os próprios autores, o projeto é bastante preliminar, pois foi o projeto de final de curso de graduação de Marcelo Gomes. Comparações de abrangência em termos de números de *synsets* foram feitas com os recursos TeP 2.0 / WN.BR, PAPEL, WN.PT, MWN.PT, e Port4Nooj. Note-se que a UfesWN.BR tem o maior número de *synsets* e a segunda maior coleção de relações entre os bancos léxicos comparados. Mas somente 31.6% dos elementos dos *synsets* da WN.Pr foram traduzidos e essas traduções não são completamente confiáveis. Por exemplo, um dos principais problemas, a desambiguação de termos, é relegado ao algoritmo do

[17] Ver http://translate.google.com/about/intl/en_ALL/license.html
[18] Ver http://logics.emap.fgv.br:10035/repositories/wn30
[19] Ver http://compling.hss.ntu.edu.sg/omw/cgi-bin/wn-gridx.cgi?gridmode=grid
[20] Universidade Federal do Espírito Santo
[21] Ver https://sites.google.com/site/ufeswordnet/

*Google Translate*, aqui usado como "caixa preta". Dados sobre a corretude dos *synsets* propostos não existem, ainda, e um sistema de verificação manual está sendo considerado. As glosas do WN.Pr também foram traduzidas e essas podem ser úteis para outros projetos, dependendo da qualidade e da facilidade de alinhá-las com esses outros projetos.

### [5.4] *Portuguese Unified Lexical Ontology*

O PULO (Simões & Guinovart 2014), abreviatura de *Portuguese Unified Lexical Ontology*, não deve ser visto como mais uma wordnet. Pretende, sim, ser o início de um projeto conjunto de disponibilização de uma wordnet livre para a língua portuguesa, perfeitamente alinhada e disponibilizada no projeto MCR: *Multilingual Central Repository* (Gonzalez-Agirre & Rigau 2013).

O início deste projeto, em finais de 2014, consistiu na realização de algumas experiências de tradução e alinhamento entre as versões inglesa, espanhola e galega da WordNet. Para além dessas mesmas wordnets, obtidas do MCR, são usados dicionários probabilísticos de tradução (Simões & Almeida 2003), um dicionário de tradução dinâmico entre as línguas portuguesa e galega (Guinovart & Simões 2013), e o vocabulário ortográfico da língua portuguesa.

Este processo foi capaz de obter cerca de 50 mil sentidos de palavras, mas apenas cerca de 17 mil foram realmente adicionadas ao PULO. Isto deveu-se ao cariz estatístico da abordagem e à linha de corte definida. O valor de pontuação obtido para cada sentido foi devidamente armazenado na base de dados de modo a que se possa ter informação da qualidade ou relevância de cada um.

A estrutura ontológica é, neste momento, a mesma que a WN.Pr, que é partilhada pelas restantes wordnets disponíveis no projeto MCR: inglês, basco, galego, castelhano e catalão. Embora o facto de se usar uma estrutura ontológica semelhante, a estrutura interna da base de dados permite que seja facilmente extensível a novos conceitos.

Neste momento, o PULO está disponível em linha[22] com 17.631 sentidos, referentes a 13.709 *synsets* diferentes. Posteriormente realizou-se uma tradução automática das glosas, usando-se para isso a API do MyMemory.[23] Através da mesma interface, é possível consultar também as restantes línguas da MCR, bem como a navegar através da ontologia base.

## [6] VISÃO COMPARATIVA

Após a descrição das várias wordnets para o português, esta secção apresenta uma comparação das suas versões mais recentes, dentro do possível, através de um conjunto de tabelas onde estas wordnets são colocadas lado a lado e ainda seguidas das mesmas propriedades para a WN.Pr. Chamamos a atenção para o fato

[22] Ver http://wordnet.pt
[23] Ver http://mymemory.translated.net/

desta comparação ser superficial e não dever ser vista como mais que isso. Muitos dos indicadores são meramente quantitativos e não consideram a coerência ou a utilidade dos conteúdos.

A tabela 1 apresenta a abordagem seguida na criação e atualização de cada wordnet e a forma de disponibilização. É notório que a alternativa mais comum à criação manual de uma wordnet para o português passa pela tradução, manual (MWN.PT), automática (UfesWN.BR), numa projeção sintática (OpenWN-PT), ou ainda em triangulação (PULO). Dentro destas quatro abordagens, o PULO destaca-se por utilizar não só a WN.Pr como wordnet "pivot", mas também as wordnets do espanhol e do galego, incluídas no MCR. Ao contrário de todas as outras, a estrutura da Onto.PT é aprendida de forma completamente automática, com base na extração de relações a partir de outros recursos textuais ou de outras wordnets, e da descoberta de aglomerados (*clusters*) de sinónimos, que dão origem aos *synsets*. Entre as vantagens de uma abordagem completamente manual, encontra-se a criação de um recurso com uma correção virtual de 100%. Por outro lado, em uma abordagem automática evita-se uma grande quantidade de trabalho cansativo, além de ser possível obter um recurso de maiores dimensões em menos tempo.

Sobre a disponibilização do recurso, recorda-se que o caráter de domínio público da WN.Pr foi um dos fatores que levou ao seu sucesso. No entanto, nem todas as wordnets para o português tomaram essa opção e apenas as quatro mais recentes são de utilização completamente livre.

| **Wordnet** | **Criação** | | **Atualização** | **Utilização** |
|---|---|---|---|---|
| | ***Synsets*** | **Relações** | | |
| WN.PT | manual | manual | manual | fechada |
| WN.BR | manual | transitividade | manual? | *synsets* livres |
| MWN.PT | tradução manual? | transitividade | ? | licença paga |
| Onto.PT | ER+*clustering* | ER+*clustering* | automática | livre |
| OpenWN-PT | projeção UWN | transitividade | semi-automática | livre |
| UfesWN.BR | tradução automática | transitividade | ? | livre |
| PULO | triangulação | transitividade | semi-automática | livre |
| WN.Pr | manual | manual | manual | livre |

TABELA 1: Wordnets do português e WN.Pr, a sua abordagem de criação e disponibilização. Apresenta-se um '?' nos casos em que desconhecemos a forma de atualização da wordnet em questão.

A tabela 2 compara a dimensão das wordnets do português relativamente ao número de itens lexicais abrangidos, separados por categoria gramatical. Neste campo a Onto.PT destaca-se por incluir um número mais de três vezes superior à segunda wordnet com mais itens lexicais, a OpenWN-PT. Isto confirma que uma abordagem de construção completamente automática será aquela com maiores possibilidades de construir um recurso de grandes dimensões num curto prazo.

Igualmente importantes para a dimensão da Onto.PT, é a quantidade (atualmente seis) e o tipo de recursos explorados, que incluíram: recursos que cobrem diferentes variantes do português, e podem levar a pequenas variações ortográficas; e dicionários, que têm já uma ampla cobertura da língua. O uso de dicionários, quer de forma manual, quer automática, é comum no processo de construção de uma wordnet. A sua exploração automática será, por um lado, uma forma da wordnet ter realmente muitas palavras e diferentes sentidos, que existem e são válidos, mas em que a utilização de uma grande fatia é pouco comum e de utilidade reduzida.

| **Wordnet** | **Itens lexicais** | | | | |
|---|---|---|---|---|---|
| | Substantivo | Verbo | Adjetivo | Advérbio | Total |
| WN.PT 1.0 | 9.813 | 633 | 485 | 0 | 10.931 |
| MWN.PT v1 | 16.000 | 0 | 0 | 0 | 16.000 |
| WN.BR | 17.000 | 10.910 | 15.000 | 1.000 | 43.910 |
| Onto.PT 0.6 | 97.531 | 32.958 | 34.392 | 3.995 | 168.876 |
| OpenWN-PT | 43.996 | 3.914 | 5.422 | 1.388 | 54.720 |
| UfesWN.BR 1.0 | 20.646 | 3.769 | 9.066 | 1.498 | 34.979 |
| PULO | 10.260 | 4.032 | 3.166 | 173 | 17.631 |
| WN.Pr 3.0 | 119.034 | 11.531 | 21.538 | 4.481 | 156.584 |

TABELA 2: Número de itens lexicais abrangidos pelas wordnets do português.

A tabela 3 apresenta outros indicadores da dimensão e cobertura, nomeadamente o número de sentidos de palavras, o número de *synsets* e ainda o número de instâncias de relações, sempre que foi possível apurar. Mais uma vez, a Onto.PT destaca-se dos demais. Olhando apenas ao número de relações, a UfesWN.BR tem um número intermédio.

É importante notar que existe um balanço intrínseco entre o número de *synsets* e a correção e utilidade da wordnet em questão. Uma das dificuldades em desenvolver uma wordnet é precisamente decidir, por um lado, se duas palavras devem ser consideradas como sinónimos, e por isso colocadas dentro do mesmo *synset* e, por outro, que palavras têm de estar em *synsets* diferentes — desafio que desde sempre acompanha lexicógrafos e para o qual, acreditamos, não há uma resposta exata. Mas parece haver um consenso de que um número muito grande de *synsets* pode ser um sinal de "ruído" no processo de agrupar palavras e/ou no processo de discriminação. Mas correção é sem dúvida um dos gargalos da construção de wordnets. Se, por um lado, dimensão e cobertura são aspectos quantitativos cuja comparação é relativamente simples (ainda que tais números, por si só, não digam muito), o mesmo não pode ser afirmado quanto à avaliação da qualidade. A WN.Pr, construída manualmente, pode até refletir decisões questionáveis, mas não contém "erros" claros, pois estamos usando-a como base de comparação. Já para as wordnets construídas de maneira automática ou semi-automática (e para

línguas que não o inglês), a avaliação da qualidade será sempre uma questão complexa, já que não há uma wordnet dourada de referência — e é justamente isso o que se quer construir. Por essa perspectiva, recursos que fazem uso do trabalho humano apresentam uma vantagem, ainda que não saibamos exatamente como esta possa ser medida.

Para as wordnets alinhadas com uma wordnet para outra língua, as relações entre *synsets* podem ser obtidas indiretamente da wordnet "pivot", por via de transitividade. Isso acontece com a MWN.PT, a OpenWN-PT, a UfesWN.BR e com o PULO. Para a WN.BR, o número de relações apresentado é apenas relativo às relações disponibilizadas juntamente com o TeP, todas elas de antonímia. Para se compreender melhor a origem destas relações, foi adicionada à tabela 3 a indicação acerca da existência de algum tipo de alinhamento com outra wordnet. Devido à sua abordagem de criação, só a Onto.PT não estará alinhada com a WN.Pr. Relativamente à WN.PT e à WN.BR sabemos que, pelo menos, houve intenções de definir um alinhamento com a WN.Pr, ainda que estes não se encontrem disponíveis — por várias vezes os autores da WN.PT mencionam o seu desenvolvimento dentro da plataforma da EuroWordNet, e os autores da WN.BR indicam como planos futuros o alinhamento da sua wordnet na mesma plataforma (Dias-da-Silva 2006).

Um alinhamento deste tipo pode ser importante para a obtenção de conhecimento adicional, a partir não só da WN.Pr, mas também de outras a ela alinhadas, o que pode ser relevante em processamento multilíngue. Para além da herança de relações, um alinhamento permite aceder a conhecimento de outras extensões da WN.Pr, tais como a WordNet-domains (Magnini & Cavaglià 2000), a SentiWordNet (Baccianella et al. 2010) ou a TempoWordNet (Dias et al. 2014), bem como alinhamentos com outros recursos, alguns dos quais referidos na secção [2.2]. Por outro lado, um alinhamento cego pode apresentar limitações relativas à cobertura na língua alvo, além de não considerar que línguas diferentes representam diferentes realidades socio-culturais, não cobrem a mesma parte do léxico e, mesmo onde parecem ser comuns, há vários conceitos lexicalizados de forma diferente (Hirst 2004). Veja-se, por exemplo, os problemas referidos na descrição do MWN.PT.

Por fim, na tabela 4 procuramos listar um conjunto de relações semânticas e indicar quais estão presentes em cada wordnet. Apesar de algumas wordnets distinguirem entre vários subtipos destas relações, optámos por utilizar uma comparação meramente booleana, em que não foi contabilizado nem o número de subtipos de cada relação, nem o número de instâncias de cada tipo. Verifica-se que apenas WN.PT e Onto.PT cobrem todas as relações listadas. No caso da Onto.PT, o conjunto de relações foi baseado no PAPEL que, por sua vez, se baseou em regularidades presentes em definições de dicionário. Alguns nomes de relação foram mesmo criados especificamente para um tipo de regularidades, o que torna

| Wordnet | Sentidos (de palavra) | *Synsets* | Relações (instâncias) | Alinhamento |
|---|---|---|---|---|
| WN.PT 1.5 | ? | 12.630 | 40.000+ | WN.Pr? |
| MWN.PT v1 | 21.000 | 17.200 | 68.735 | WN.Pr |
| WN.BR | 75,720 | 19.888 | 4.276+? | WN.Pr? |
| Onto.PT 0.6 | 248.773 | 117.450 | 341.506 | nenhum |
| OpenWN-PT | 73.802 | 43.925 | 74.102 | WN.Pr |
| UfesWN.BR 1.0 | 63.096 | 48.981 | 238.413 | WN.Pr |
| PULO | 17.631 | 13.709 | 48.658 | MCR |
| WN.Pr | 206.978 | 117.659 | 285.000 | — |

TABELA 3: *Synsets* e relações nas wordnets do português. Indicamos com '?' casos em que não conseguirmos confirmar a informação.

o conjunto bastante rico. Igualmente rico é o conjunto da WN.PT, que não só cobre todas as relações listadas como, de acordo com diferentes especificidades de cada relação, tem vários subtipos de fundamentação linguística. A diferença do conjunto de relações da WN.PT com a WN.Pr levanta dúvidas acerca do tipo de alinhamento entre estes dois recursos. Para as demais wordnets, dado o seu alinhamento, o conjunto de relações coberto é o mesmo que o da WN.Pr.

| Wordnets | Relações | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | Sinon | Anton | Hiperon | Meron | Causa | Finalid | Local | Maneira |
| WN.PT | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| MWN.PT | ✓ | × | ✓ | ✓ | × | × | × | × |
| WN.BR | ✓ | ✓ | ✓ | ✓ | ✓ | × | × | × |
| Onto.PT | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| OpenWN-PT | ✓ | ✓ | ✓ | ✓ | ✓ | × | × | × |
| UfesWN.BR | ✓ | ✓ | ✓ | ✓ | ✓ | × | × | × |
| PULO | ✓ | ✓ | ✓ | ✓ | ✓ | × | × | × |
| WN.Pr | ✓ | ✓ | ✓ | ✓ | ✓ | × | × | × |

TABELA 4: Relações semânticas nas wordnets do português.

## [7] DISCUSSÃO FINAL

Foram apresentadas e, dentro do possível, comparadas as várias wordnets que existem atualmente para a língua portuguesa. Entre elas, há quatro wordnets e uma base de *synsets* (TeP/WN.BR) livremente disponíveis, para além de uma wordnet que pode ser comprada (MWN.PT) e de outra que pode ser explorada em linha (WN.PT). A construção destas wordnets seguiu abordagens diferentes, desde trabalho completamente manual, passando por abordagens baseadas em tradução, com mais ou menos trabalho manual, ou ainda uma abordagem em que toda a estrutura da wordnet é aprendida de forma automática. Esperamos ter mostrado que, atualmente, já não faz sentido lamentar que não existe uma wordnet

para o português. Aliás, a utilização de uma wordnet num projeto que vise a língua portuguesa é cada vez menos um problema com uma solução de remedeio, e cada vez mais um problema de escolha dentro das alternativas disponíveis. Esta escolha deverá considerar, entre outros, a necessidade de alinhamento com outras wordnets, a tolerância a erros, a necessidade de abrangência — tanto no que diz respeito às relações presentes quanto aos itens lexicais cobertos — ou mesmo o orçamento disponível. Uma vez que cada wordnet tem caraterísticas diferentes das demais, também não será de descartar a utilização de mais de uma no mesmo projeto.

Será também pertinente perguntar se esta quantidade de alternativas faz sentido ou se seria preferível os autores interessados focarem-se na construção de uma única wordnet para o português, tentando aproveitar as forças de cada um dos projetos descritos.

Os autores deste artigo, responsáveis pela Onto.PT, OpenWN-PT e PULO, acreditam que haverá vantagens nas duas opções e, por isso, nos próximos tempos, será seguida uma abordagem intermédia. Ou seja, o desenvolvimento de cada wordnet continuará a ser feito pelas mesmas equipas que o têm feito até aqui, mas haverá um maior acompanhamento do trabalho desenvolvido por cada equipa. Desta forma, entre outras vantagens, cada projeto poderá tirar partido do que é feito nos outros, minimizando a quantidade de trabalho duplicado, mas sem perder de vista objetivos específicos de cada projeto.

Como seria de esperar, é comum aos três projetos a vontade de continuar a melhorar a coerência, qualidade e abrangência do seu recurso. Para além de tarefas já planeadas, a médio e a longo prazo, específicas para cada um dos projeto, os autores deste artigo vêem com bom olhos futuras colaborações que possam tirar partido do que já foi feito e que, a longo prazo, possam até levar a uma integração ou alinhamento dos seus projetos. Assim, após enumerar os objetivos individuais, serão indicadas linhas gerais de potenciais colaborações que surgem no seguimento de algumas discussões entre os autores.

### [7.1] *Trabalho futuro independente*

A Onto.PT deverá continuar a ser construída com base na extração automática de informação a partir de recursos lexicais disponíveis para o português. Juntamente com os recursos atualmente explorados, será considerada a utilização de outros, tais como as definições ou os triplos semânticos do Port4Nooj, a Wikipédia ou os conteúdos das restantes wordnets. Continuarão a ser feitas avaliações dos vários métodos automáticos, de forma a conseguir melhorar a qualidade do recurso, e poderão ser feitas avaliações manuais pontuais, que possam indicar o progresso dessa mesma qualidade. Uma evolução estrutural da Onto.PT será a associação de um valor numérico, indicador da confiança nos seus conteúdos e, eventualmente, outro valor indicador da frequência de utilização das palavras em corpos. Esta

adição daria uma nova dimensão ao recurso e permitiria um tipo de utilização diferente. Ou seja, a Onto.PT poderá continuar a crescer, mas os seus utilizadores poderão definir um limite na confiança dos conteúdos a usar. Deve ainda dizer-se que, dada a sua abordagem de construção, a Onto.PT não é estática, nem no tocante a palavras, nem a *synsets* ou relações. Assim, ainda que seja possível esboçar algo intermédio, não está nos planos um futuro alinhamento com a WN.Pr.

As prioridades da OpenWN-PT incluem a uniformização de conteúdos, bem como uma verificação da abrangência dos *synsets* nucleares da GWA. A uniformização passará, entre outros, por verificar se as classes morfológicas dos termos estão corretas, por lematizar termos não lematizados, e por garantir que exista uma coleção compreensiva de verbos, agrupados e classificados de acordo com suas *frames* de subcategorização. A discussão de alguns problemas encontrados e possíveis soluções foi já começada (de Paiva et al. 2014a). Outro objetivo da OpenWN-PT é aprimorar a interface de busca,[24] validação e atualização, para que isso facilite o trabalho colaborativo de melhoria contínua do recurso. Paralelamente, existe a intenção de expandir o recurso baseando-se em corpos ou em outros recursos, como o PAPEL. Pretende-se fazê-lo não só ao nível da introdução de novas palavras/sentidos, mas também ao nível de novas relações e glosas para os *synsets*, possivelmente traduzidas da WN.Pr e complementadas com definições de dicionários abertos. A OpenWN-PT pretende manter o desenvolvimento em conjunto com o NomLex-PT, e manter a ligação com a GWA e com a OMWN. A longo prazo, os seus autores pretendem também aprimorar o mapeamento da OpenWN-PT para a ontologia SUMO (obtido através do mapeamento de WN.Pr), o que vai permitir o estabelecimento de um sistema de raciocínio lógico sobre conhecimentos obtidos através de PLN, à semelhança do sistema BRIDGE, desenvolvido pelo grupo da Xerox PARC (Bobrow et al. 2007).

O PULO, como o projeto mais jovem dos três, pretende, para já, consolidar a sua integração no MCR. Além disso, está a ser desenvolvida uma interface de validação dos conteúdos através da inteligência das massas (vulgo *crowdsourcing*). A longo prazo, há também a intenção de alinhar o PULO com outros projetos dos seus autores, nomeadamente o Dicionário Aberto.

### [7.2] *Linhas para trabalho conjunto*

Há várias ideias para futuras colaborações entre os projetos, muitas delas já discutidas pelos autores, e que passamos a enumerar:

(i) A wordnet Onto.PT é apenas um das contribuições do projeto Onto.PT. Talvez a mais importante tenha até sido a abordagem ECO, adotada na sua construção, e que visa não só criar wordnets de raiz, mas que pode também ser usada para enriquecer wordnets existentes. Assim, fará todo o sentido apli-

[24] Ver http://logics.emap.fgv.br/wn/

car alguns dos seus procedimentos automáticos para sugerir novos conteúdos quer à OpenWN-PT, quer ao PULO, nomeadamente: (i) novas palavras a *synsets*; (ii) novas instâncias de relações abrangidas; (iii) novos tipos de relação; (iv) glosas.

(ii) Há já interfaces de busca para as três wordnets dos autores, criadas pelos próprios autores ou por terceiros. No entanto, OpenWN-PT e PULO querem ir mais longe e ter uma interface de sugestão e validação dos conteúdos. Tanto OpenWN-PT como PULO têm já protótipos para essa interface, e o seu desenvolvimento poderia ser feito em parceria.

(iii) Seria interessante fazer uma ponte entre outros recursos desenvolvidos pelos mesmos autores. Isto incluiria, por exemplo, um alinhamento do NomLex-PT com o PULO e o Dicionário Aberto não só com PULO, mas também o OpenWN-PT. O Dicionário Aberto poderia mesmo ser utilizado como uma fonte adicional de glosas em português para os *synsets* de qualquer uma das wordnets.

(iv) Os conteúdos de OpenWN-PT e PULO poderão ser explorados pela Onto.PT e ser uma fonte adicional para calcular o tal valor numérico de confiança. Aliás, à medida que estes recursos forem atingindo um nível maior de coerência, poderão também vir a ser usados como referência na avaliação da Onto.PT.

De modo a perceber até que ponto estas wordnets se complementam ou não, e até que ponto faria sentido e seria possível algum tipo de integração ou alinhamento, o ponto de partida para uma colaboração mais estreita deveria passar por uma comparação mais exaustiva das três, incluindo, dentro do possível, as restantes wordnets livres. No que diz respeito às wordnets que estão alinhadas com a WN.Pr, a comparação será provavelmente mais fácil e direta.

Por sua vez, a comparação poderia começar de uma forma muito simples, com a criação de uma ligação na interface de cada wordnet que permitisse obter os resultados da mesma pesquisa nas demais wordnets. Poderia depois passar por selecionar aleatoriamente um conjunto de palavras e analisar não só a sua presença nas várias wordnets, como os seus sentidos. Mas dada a dificuldade em avaliar diretamente uma wordnet, um possível atalho envolveria a seleção de frases padrão teste, em linguagem natural, que transmitam determinada relação semântica de forma objetiva (por exemplo, *<x> é um tipo de <y>*, para hiperonímia, ou *<x> tem um <y>*, para meronímia). A partir dessas frases, podem ser geradas variantes através da substituição das duas palavras relacionadas pelos argumentos de qualquer relação do mesmo tipo. Um avaliador deverá depois indicar se cada frase resultado mantém a coerência semântica. Isto foi já proposto por (Cruse 1986) e

seguido, entre outros, na avaliação da WN.PT (Marrafa 2002) e numa avaliação inicial da OpenWN-PT (Rademaker et al. 2014). Poderia ainda recorrer-se ao sistema VARRA (Freitas et al. 2014b) para, por exemplo, procurar por frases de corpos em que pares de palavras relacionadas co-ocorram.

Por último, a utilidade das wordnets poderia ser também medida, por exemplo, através do seu desempenho num conjunto de tarefas de PLN bem definidas, onde fosse necessário recorrer precisamente a uma wordnet. Entre outras tarefas, destacam-se a recuperação de informação, tarefa para a qual a Onto.PT já foi usada (Rodrigues et al. 2012), ou a resposta automática a questões de escolha múltipla, onde uma comparação entre diferentes recursos léxico-semânticos foi recentemente ensaiada (Gonçalo Oliveira et al. 2014).

## AGRADECIMENTOS

Um agradecimento à Belinda Maia, co-organizadora do *Workshop on Language Resources for Teaching and Research* e da Escola de Verão da Linguateca, ambos realizados na Faculdade de Letras da Universidade do Porto, onde o primeiro autor deste artigo (Hugo) teve o prazer de ser convidado a apresentar o seu trabalho desenvolvido no âmbito do PAPEL, que viria a dar origem à sua investigação em torno da construção de wordnets.

## REFERÊNCIAS

Amaro, Raquel. 2014. Extracting semantic relations from portuguese corpora using lexical-syntactic patterns. Em *Proceedings of the 9th international conference on language resources and evaluation* LREC'14, ELRA.

van Assem, Mark, Aldo Gangemi & Guus Schreiber. 2006. RDF/OWL representation of WordNet. W3C working draft World Wide Web Consortium. http://www.w3.org/TR/2006/WD-wordnet-rdf-20060619/.

Baccianella, Stefano, Andrea Esuli & Fabrizio Sebastiani. 2010. SentiWordNet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. Em *Proceedings of 7th International Conference on Language Resources and Evaluation*, 2200–2204. ELRA.

Banerjee, Satanjeev & Ted Pedersen. 2002. An adapted Lesk algorithm for word sense disambiguation using WordNet. Em *Proceedings of the 3rd international conference on computational linguistics and intelligent text processing (CICLing 2002)*, vol. 2276 LNCS, 136–145. Springer.

Barreiro, Anabela. 2010. Port4NooJ: an open source, ontology-driven portuguese linguistic system with applications in machine translation. Em *Proceedings of the 2008 international nooj conference (nooj'08)*, Cambridge Scholars Publishing.

Barreiro, Anabela, Fernando Batista, Ricardo Ribeiro, Helena Moniz & Isabel Trancoso. 2014. OpenLogos Semantico-Syntactic Knowledge-Rich Bilingual Dictionaries. Em Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk & Stelios Piperidis (eds.), *Proceedings of the Ninth International Conference on Language Resources and Evaluation*, 3774–3781. ELRA.

Bobrow, Daniel G, Bob Cheslow, Cleo Condoravdi, Lauri Karttunen, Tracy Holloway King, Rowan Nairn, Valeria de Paiva, Charlotte Price & Annie Zaenen. 2007. Parc's bridge and question answering system. Em Tracy Holloway King & Emily M. Bender (eds.), *Proceedings of the geaf 2007 workshop.*, 46–66. CSLI.

Bond, Francis & Ryan Foster. 2013. Linking and extending an open multilingual wordnet. Em *Proceedings of the 51st annual meeting of the association for computational linguistics*, vol. 1, 1352–1362. ACL.

Bond, Francis & Kyonghee Paik. 2012. A survey of wordnets and their licenses. Em *Proceedings of the 6th global wordnet conference*, 64–71.

Cruse, Alan D. 1986. *Lexical semantics*. Cambridge University Press.

Dias, Gaël Harry, Mohammed Hasanuzzaman, Stéphane Ferrari & Yann Mathet. 2014. TempoWordNet for Sentence Time Tagging. Em *Proceedings of the Companion Publication of the 23rd International Conference on World Wide Web Companion*, 833–838.

Dias-da-Silva, Bento C. 2006. Wordnet.Br: An exercise of human language technology research. Em *Proceedings of 3rd international wordnet conference (gwc)*, 301–303.

Dias-da-Silva, Bento C., Mirna F. de Oliveira & Helio R. de Moraes. 2002. Groundwork for the Development of the Brazilian Portuguese Wordnet. Em *Advances in Natural Language Processing (PorTAL 2002)*, 189–196. Springer.

Drury, Brett, Paula C.F. Cardoso, Janie M. Thomas & Alneu de Andrade Lopes. 2014. Lexical resources for the identification of causative relations in portuguese texts. Em *Proceedings of the 1st Workshop on Tools and Resources for Automatically Processing Portuguese and Spanish*, 56–63.

Fellbaum, Christiane (ed.). 1998. *WordNet: An Electronic Lexical Database (language, speech, and communication)*. The MIT Press.

Fellbaum, Christiane. 2010. WordNet. Em *Theory and applications of ontology: Computer applications*, chap. 10, 231–243. Springer.

Finlayson, Mark. 2014. Java libraries for accessing the princeton wordnet: Comparison and evaluation. Em *Proceedings of the Seventh Global Wordnet Conference (GWC)*, 78–85.

Freitas, Cláudia, Valeria de Paiva, Alexandre Rademaker, Gerard de Melo, Livy Real & Anne de Araujo Correia da Silva. 2014a. Extending a lexicon of portuguese nominalizations with data from corpora. Em Jorge Baptista, Nuno Mamede, Sara Candeias, Ivandré Paraboni, Thiago A. S. Pardo & Maria das Graças Volpe Nunes (eds.), *Computational Processing of the Portuguese Language, 11th International Conference (PROPOR)*, Springer.

Freitas, Cláudia, Diana Santos, Hugo Gonçalo Oliveira & Violeta Quental. 2014b. VARRA: Validação, Avaliação e Revisão de Relações semânticas no AC/DC. Em *Pesquisas e perspetivas em linguistica de corpus*, Mercado de Letras.

Gangemi, Aldo, Nicola Guarino, Claudio Masolo & Alessandro Oltramari. 2010. Interfacing WordNet with DOLCE: towards OntoWordNet. Em *Ontology and the lexicon: A natural language processing perspective* Studies in Natural Language Processing, chap. 3, 36–52. Cambridge University Press.

George A. Miller and Martin Chodorow and Shari Landes and Claudia Leacock and Robert G. Thomas. 1994. Using a semantic concordance for sense identification. Em *Proceedings of ARPA Human Language Technology Workshop*, 240–243.

Gomes, Marcelo Machado, Walber Beltrame & Davidson Cury. 2013. Automatic Construction of Brazilian Portuguese WordNet. Em *Proceedings of X National Meeting on Artificial and Computational Intelligence*, s/pp.

Gonçalo Oliveira, Hugo. 2013. *Onto.PT: Towards the Automatic Construction of a Lexical Ontology for Portuguese*: University of Coimbra. Tese de Doutoramento.

Gonçalo Oliveira, Hugo. 2014. On the utility of Portuguese term-based lexical-semantic networks. Em *Proceedings of Computational Processing of the Portuguese Language - 11th International Conference (PROPOR)*, vol. 8775, 176–182. Springer.

Gonçalo Oliveira, Hugo, Inês Coelho & Paulo Gomes. 2014. Exploiting Portuguese lexical knowledge bases for answering open domain cloze questions automatically. Em *Proceedings of the 9th Language Resources and Evaluation Conference (LREC)*, ELRA.

Gonçalo Oliveira, Hugo & Paulo Gomes. 2010. Onto.PT: Automatic Construction of a Lexical Ontology for Portuguese. Em *Proceedings of 5th European Starting AI Researcher Symposium (STAIRS 2010)*, vol. 222, 199–211. IOS Press.

Gonçalo Oliveira, Hugo & Paulo Gomes. 2014a. ECO and Onto.PT: A flexible approach for creating a Portuguese wordnet automatically. *Language Resources and Evaluation* 48(2). 373–393.

Gonçalo Oliveira, Hugo & Paulo Gomes. 2014b. Onto.PT: recent developments of a large public domain portuguese wordnet. Em *Proceedings of the 7th Global WordNet Conference*, 16–22.

Gonçalo Oliveira, Hugo, Leticia Antón Pérez, Hernani Costa & Paulo Gomes. 2011. Uma rede léxico-semântica de grandes dimensões para o português, extraída a partir de dicionários electrónicos. *Linguamática* 3(2). 23–38.

Gonçalo Oliveira, Hugo, Diana Santos, Paulo Gomes & Nuno Seco. 2008. PAPEL: A dictionary-based lexical ontology for Portuguese. Em *Proceedings of Computational Processing of the Portuguese Language - 8th International Conference (PROPOR)*, vol. 5190, 31–40. Springer.

Gonzalez-Agirre, Aitor & German Rigau. 2013. Construcción de una base de conocimiento léxico multilingüe de amplia cobertura: Multilingual Central Repository. *Linguamática* 5(1). 13–28.

Guinovart, Xavier Gómez & Alberto Simões. 2013. Retreading Dictionaries for the 21st Century. Em José Paulo Leal, Ricardo Rocha & Alberto Simões (eds.), *2nd Symposium on Languages, Applications and Technologies*, vol. 29, 115–126. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik.

Gurevych, Iryna, Judith Eckle-Kohler, Silvana Hartmann, Michael Matuschek, Christian M. Meyer & Christian Wirth. 2012. UBY - a large-scale unified lexical-semantic resource. Em *Proceedings of the 13th conference of the european chapter of the association for computational linguistics*, 580–590. ACL Press.

Hirst, Graeme. 2004. Ontology and the lexicon. Em Steffen Staab & Rudi Studer (eds.), *Handbook on ontologies* International Handbooks on Information Systems, 209–230. Springer.

Kilgarriff, Adam. 1997. I don't believe in word senses. *Computers and the Humanities* 31. 91–113.

Magnini, Bernardo & Gabriela Cavaglià. 2000. Integrating subject field codes into WordNet. Em *Proceedings of 2nd International Conference on Language Resources and Evaluation (LREC)*, 1413–1418. ELRA.

Marrafa, Palmira. 2001. *Wordnet do português: uma base de dados de conhecimento linguístico*. Instituto Camões.

Marrafa, Palmira. 2002. Portuguese WordNet: general architecture and internal semantic relations. *DELTA* 18. 131–146.

Marrafa, Palmira, Raquel Amaro & Sara Mendes. 2011. WordNet.PT Global – extending WordNet.PT to Portuguese varieties. Em *Proceedings of 1st workshop on algorithms and resources for modelling of dialects and language varieties*, 70–74. ACL Press.

Maziero, Erick G., Thiago A. S. Pardo, Ariani Di Felippo & Bento C. Dias-da-Silva. 2008. A Base de Dados Lexical e a Interface Web do TeP 2.0: Thesaurus Eletrônico para o Português do Brasil. Em *VI Workshop em Tecnologia da Informação e da Linguagem Humana*, 390–392.

de Melo, Gerard & Gerhard Weikum. 2009. Towards a universal wordnet by learning from combined evidence. Em *Proceedings of the 18th ACM Conference on Information and Knowledge Management (CIKM)*, 513–522. ACM.

de Melo, Gerard & Gerhard Weikum. 2012. UWN: A large multilingual lexical knowledge base. Em *Proceedings of the 50th annual meeting of the association for computational linguistics*, 151–156. Stroudsburg, PA, USA: Association for Computational Linguistics.

Miller, George A. 1995. Wordnet: a lexical database for english. *Communications of the ACM* 38(11). 39–41.

Naber, Daniel. 2004. Openthesaurus: Building a thesaurus with a Web community. (retrieved on August 2012). http://www.openthesaurus.de/download/openthesaurus.pdf.

Navigli, Roberto & Simone Paolo Ponzetto. 2012. BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence* 193. 217–250.

Padró, Lluís & Evgeny Stanilovsky. 2012. Freeling 3.0: Towards wider multilinguality. Em *Proceedings of the Language Resources and Evaluation Conference (LREC)*, 2473–2479.

de Paiva, Valeria, Cláudia Freitas, Livy Real & Alexandre Rademaker. 2014a. Improving the Verb Lexicon of OpenWordnet-PT. Em Laura Alonso Alemany, Muntsa Padró, Alexandre Rademaker & Aline Villavicencio (eds.), *Proceedings of Workshop on Tools and Resources for Automatically Processing Portuguese and Spanish*, 110–115.

de Paiva, Valeria, Alexandre Rademaker & Gerard de Melo. 2012. OpenWordNet-PT: An open brazilian wordnet for reasoning. Em *Proceedings of 24th International Conference on Computational Linguistics*, 353–360.

de Paiva, Valeria, Livy Real, Alexandre Rademaker & Gerard de Melo. 2014b. NomLex-PT: A Lexicon of Portuguese Nominalizations. Em *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC)*, 114–124. ELRA.

Pease, Adam & Christiane Fellbaum. 2010. Formal ontology as interlingua: the SUMO and WordNet linking project and global WordNet linking project. Em *Ontology and the Lexicon: A Natural Language Processing Perspective*, chap. 2, 25–35. Cambridge University Press.

Pianta, Emanuele, Luisa Bentivogli & Christian Girardi. 2002. MultiWordNet: developing an aligned multilingual database. Em *Proceedings of 1st International Conference on Global WordNet*, 293–302.

Rademaker, Alexandre, Valeria De Paiva, Gerard de Melo, Livy Maria Real Coelho & Maira Gatti. 2014. OpenWordNet-PT: A Project Report. Em *Proceedings of the 7th Global WordNet Conference*, 383–390.

Resnik, Philip. 1995. Using information content to evaluate semantic similarity in a taxonomy. Em *Proceedings of the 14th International Joint Conference on Artificial Intelligence*, 448–453. Morgan Kaufmann.

Rodrigues, Ricardo, Hugo Gonçalo Oliveira & Paulo Gomes. 2012. Uma abordagem ao Págico baseada no processamento e análise de sintagmas dos tópicos. *Linguamática* 4(1). 31–39.

Sampson, Geoffrey. 2000. Review of Fellbaum (1998). *International Journal of Lexicography* 13(1). 54–59.

Santos, Diana, Anabela Barreiro, Cláudia Freitas, Hugo Gonçalo Oliveira, José Carlos Medeiros, Luís Costa, Paulo Gomes & Rosário Silva. 2010. Relações semânticas em português: comparando o TeP, o MWN.PT, o Port4NooJ e o PAPEL. Em *Textos seleccionados. XXV Encontro Nacional da Associação Portuguesa de Linguística*, 681–700.

Silberztein, Max. 2005. NooJ: A Linguistic Annotation System for Corpus Processing. Em *Proceedings of HLT/EMNLP on Interactive Demonstrations*, 10–11. ACL Press.

Simões, Alberto & Xavier Gómez Guinovart. 2014. Bootstrapping a Portuguese wordnet from Galician, Spanish and English wordnets. Em *Advances in Speech and Language Technologies for Iberian Languages*, vol. 8854, 239–248. Springer.

Simões, Alberto, Álvaro Iriarte Sanromán & José João Almeida. 2012. Dicionário-Aberto: A source of resources for the Portuguese language processing. Em *Proceedings of Computational Processing of the Portuguese Language, 10th International Conference (PROPOR)*, vol. 7243, 121–127. Springer.

Simões, Alberto M. & J. João Almeida. 2003. NATools – a statistical word aligner workbench. *Procesamiento del Lenguaje Natural* 31. 217–224.

Stamou, Sofia, Kemal Oflazer, Karel Pala, Dimitris Christoudoulakis, Dan Cristea, Dan Tufis, Svetla Koeva, George Totkov, Dominique Dutoit & Maria Grigoriadou. 2002. BalkaNet: A multilingual semantic network for the balkan languages. Em *Proceedings of 1st Global WordNet Conference*, 3–4.

Suchanek, Fabian M., Gjergji Kasneci & Gerhard Weikum. 2007. YAGO: a core of semantic knowledge. Em *Proceedings of the 16th international conference on world wide web*, 697–706. ACM.

Vossen, Piek. 1997. EuroWordNet: a multilingual database for information retrieval. Em *Proceedings of DELOS workshop on cross-language information retrieval*, 5–7.

## CONTACTOS

Hugo Gonçalo Oliveira
CISUC, Universidade de Coimbra, Portugal
hroliv@dei.uc.pt

Valeria de Paiva
Nuance Communications, USA
valeria.depaiva@nuance.com

Cláudia Freitas
PUC-Rio, Brasil
claudiafreitas@puc-rio.br

Alexandre Rademaker
IBM Research e FGV/EMAp, Brasil
alexrad@br.ibm.com

Livy Real
IBM Research, Brasil
livyreal@gmail.com

Alberto Simões
CEH, Universidade do Minho e Linguateca
ambs@ilch.uminho.pt